IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 227–240

DOI: 10.1214/12-IMSCOLL916

Analyzing posteriors by the information inequality

Willem Kruijer and Aad van der Vaart

Wageningen UR and Universiteit Leiden

Abstract: We give bounds on the concentration of (pseudo) posterior distributions, both for correct and misspecified models. The bounds are derived using the information inequality, entropy estimates, and empirical process methods.

1. Introduction

The *posterior distribution* corresponding to a prior probability distribution Π on a set $\mathcal{P}$ of probability densities on a given measurable space $(\mathcal{X}, \mathcal{A})$ is the random probability measure defined through

$$d\Pi(p|\, X) \propto p(X)\, d\Pi(p). \tag{1}$$

Here the element X of $\mathcal{X}$ is considered distributed according to some fixed true density q on $(\mathcal{X}, \mathcal{A})$, which may or may not belong to $\mathcal{P}$. To make the expression well defined we assume that Π is a probability distribution on a σ-field on $\mathcal{P}$ for which the map $(x, p) \mapsto p(x)$ is jointly measurable, that the dominating measure μ for $\mathcal{P}$ on $(\mathcal{X}, \mathcal{A})$ is σ-finite, and that the "norming constant" $\int p(X)\, d\Pi(p)$ is finite and positive with probability one under q.

Several authors have studied whether the posterior distribution can recover the true density q, often in an asymptotic setting where X is a vector of n i.i.d. observations and $n \to \infty$. The study of *posterior consistency*, the contraction of a sequence of posterior distributions to a Dirac measure at q, was initiated by [9], while study of the rate of contraction, in the nonparametric situation, was taken up more recently by [2]. These papers phrase their results in terms of a testing criterion, which can be traced back to [8]. Subsequently refinements and different approaches were found. In the present note we give a simplified presentation of the interesting approach by [13], which is based on the information inequality, and relate it to the testing approach. We also cover misspecified models and the range of pseudo posteriors that bridge the gap between Bayes and maximum likelihood.

We are mainly interested in the true posterior distribution (1), but consider, more generally, the random probability measures defined by, for $\rho > 0$,

$$d\Pi_\rho(p|\, X) \propto p^\rho(X)\, d\Pi(p). \tag{2}$$

For $\rho \in (0, 1]$ these distributions are defined as soon as the posterior distribution, which is the special case $\rho = 1$, is defined. For $\rho > 1$ finiteness of the norming integral $\int p^\rho(X)\, d\Pi(p)$ is not automatic, but must be assumed.

P.O. Box 9512, 2300 RA Leiden, e-mail: avdvaart@math.leidenuniv.nl, url: http://www.math.leidenuniv.nl/∼avdvaart

AMS 2000 subject classifications: Primary 60K35
Keywords and phrases: posterior contraction, prior, Bayes

It turns out that results are easiest to obtain for the random measures with $\rho < 1$. This makes this choice attractive for the purpose of recovery of a true parameter. The disadvantage is that these "pseudo-posteriors" lack a clear interpretation, which may also make them computationally inaccessible. Admittedly not much is known at this time about the frequentist meaning of the spread in the (pseudo) posterior distribution (and the corresponding *posterior credibility sets*), so that even the interpretation of the true posterior distribution may not extend beyond the Bayesian realm.

For increasing ρ the "pseudo likelihood" $p \mapsto p^\rho(X)$ increasingly accentuates the high points of the likelihood and decreases its lows. The pseudo posterior in the limit case $\rho = \infty$ could be interpreted as a Dirac measure at the maximum likelihood estimator(s). The potential instability of the nonparametric maximum likelihood estimator and stability of a Bayesian estimator is well documented. It seems interesting that further deaccentuating the heights in the likelihood ($\rho < 1$) increases the stability.

We note that "stability" means here that the method works in more situations. It is not a measure of quality in a given situation, when multiple methods work.

2. Information theory

For nonnegative, integrable functions p and q on a measure space $(\mathcal{X}, \mathcal{A}, \mu)$, and $\alpha > 0$, we define

$$\begin{aligned}
\rho_\alpha(p,q) &= \int p^\alpha q^{1-\alpha}\, d\mu && \textit{(Hellinger transform)},\\
R_\alpha(p,q) &= -\log \int p^\alpha q^{1-\alpha}\, d\mu && \textit{(Renyi divergence)},\\
KL(p,q) &= \int \big(\log(q/p)\big) q\, d\mu && \textit{(Kullback-Leibler divergence)},\\
h(p,q) &= \sqrt{\int (\sqrt{p}-\sqrt{q})^2\, d\mu} && \textit{(Hellinger distance)}.
\end{aligned}$$

For $\alpha > 1$ the Hellinger transform and negative Renyi divergence may be infinite, depending on p and q. The Kullback-Leibler divergence may be infinite, but is always well defined; by convention $K(p,q) = \infty$ if $Q(p = 0) > 0$. We note that the Hellinger distance is sometimes defined as our $h(p,q)/2$; furthermore, the order of the arguments in $K(p,q)$ may differ.

In the following lemma we recall some elementary properties. Let P denote the measure with density p, and let $\|P\| = P(\mathcal{X}) = \int p\, d\mu$ denote its L_1-norm.

Lemma 2.1. *For nonnegative integrable functions p and q the map $\alpha \mapsto \rho_\alpha(p,q)$ is convex on $[0,1]$ with limits $Q(p > 0)$ and $P(q > 0)$, and derivatives $-KL(p, q1_{p>0})$ and $-KL(q, p1_{q>0})$ at $\alpha = 0$ at $\alpha = 1$. Furthermore, the maps $p \mapsto \rho_\alpha(p,q)$ and $p \mapsto KL(p,q)$ from $L_1(\mu)$ to $\mathbb{R}$ are upper and lower semicontinuous, respectively, and for $\alpha \in (0,1)$,*

(i) $\rho_\alpha(p,q) \le \|P\|^\alpha \, \|Q\|^{1-\alpha} \le \alpha\|P\| + (1-\alpha)\|Q\|$.
(ii) $h^2(p,q) = \|P\| + \|Q\| - 2\rho_{1/2}(p,q)$.
(iii) $(\alpha \wedge (1-\alpha))h^2(p,q) \le \alpha\|P\| + (1-\alpha)\|Q\| - \rho_\alpha(p,q) \le h^2(p,q)$.
(iv) $\|Q\| - \rho_\alpha(p,q) \le \alpha KL(p,q)$, *if* $Q \ll P$.
(v) $h^2(p,q) + \|Q\| - \|P\| \le KL(p,q)$.

Finally, for probability densities p and q, and $\alpha \in (0,1)$,

(vi) $R_\alpha(p,q) \geq 0$.
(vii) $1 - \rho_\alpha(p,q) \leq R_\alpha(p,q) \leq \rho_\alpha^{-1}(p,q) - 1$.
(viii) $(\alpha \wedge (1-\alpha))h^2(p,q) \leq R_\alpha(p,q) \leq h^2(p,q)/(1-h^2(p,q))$, *if* $h(p,q) < 1$.
(ix) $\alpha^{-1}(1-\alpha)^{-1}R_\alpha(p,q)$ *tends to* $KL(p,q)$ *and* $KL(q,p)$ *as* $\alpha \downarrow 0$ *or* $\alpha \uparrow 1$, *respectively, if* P *and* Q *are mutually absolutely continuous.*

Proof. The first assertion follows from convexity of the map $\alpha \mapsto e^{\alpha y}$, for any $y \in \mathbb{R}$; for a precise proof see e.g. [5]. Statement (i) follows from Hölder's and Young's inequalities. The lower inequality of statement (iii) for $\alpha < 1/2$ follows from rearranging the inequality $\rho_\alpha \leq (1-2\alpha)\rho_0 + 2\alpha\rho_{1/2}$, which is a consequence of the convexity of $\alpha \mapsto \rho_\alpha$, combined with the bound (i) on ρ_0 and the rewrite (ii) of $\rho_{1/2}$; the inequality for $\alpha \geq 1/2$ follows similarly from $\rho_\alpha \leq (2-2\alpha)\rho_{1/2} + (2\alpha-1)\rho_1$. The upper inequality follows similarly from considering $1/2$ as the convex combination of α and $1-\alpha$. Assertion (iv) is equivalent to $\rho_0(p,q) - \alpha KL(p, q1_{p>0})\alpha \leq \rho_\alpha(p,q)$, which is true again by convexity and the fact that $KL(p, q1_{p>0})$ is the derivative of $\alpha \mapsto \rho_\alpha(p,q)$ at $\alpha = 0$. Statement (v) follows from combining (iv) (with $\alpha = 1/2$) and (ii) if $Q \ll P$; in the other case it is trivial, because $KL(p,q) = \infty$. Assertion (vii) follows from $1 - x \leq -\log x \leq 1/x - 1$, for $x > 0$. Inequalities (viii) are found by combining (vii) with (iii). □

Part (viii) of the lemma shows that for probability densities any Renyi divergence is (almost) interchangeable with the squared Hellinger distance. An advantage of the former is its exact additivity for product measures. Unfortunately, the equivalence does not extend to general nonnegative functions. Part (iii) of the lemma suggests to redefine the Renyi divergence as $R_\alpha(p,q) + \log(\alpha\|P\| + (1-\alpha)\|Q\|)$ if p or q do not integrate to one, in which case it becomes again comparable to $h^2(p,q)$.

For probability densities the Kullback-Leibler divergence dominates the squared Hellinger distance, and hence essentially also the Renyi divergence, but by its asymmetry it does not compare easily on arguments with different total masses.

For a collection $\mathcal{P}$ of densities we define

$$\rho_\alpha(\mathcal{P}, q) = \sup_{p \in \text{conv }(\mathcal{P})} \rho_\alpha(p,q),$$
$$R_\alpha(\mathcal{P}, q) = \inf_{p \in \text{conv }(\mathcal{P})} R_\alpha(p,q),$$
$$KL(\mathcal{P}, q) = \sup_{p \in \text{conv }(\mathcal{P})} KL(p,q).$$

Here conv $(\mathcal{P})$ denotes the convex hull of $\mathcal{P}$, defined as the set of all averages $\int p\, d\Pi(p)$ relative to priors Π on $\mathcal{P}$. One motivation for taking the supremum or infimum over the convex hull is that the functionals become sub-multiplicative and super-additive relative to product measures. See Lemma 4.1. Because the Kullback-Leibler divergence is convex in its arguments, taking the supremum over the convex hull rather than over just $\mathcal{P}$ does not make the expression bigger in this case.

The Hellinger transform, as a function of α, is well known from the theory of statistical experiments (see [7]). The function $\alpha \mapsto \rho_\alpha(p,q)$ fully characterizes the binary statistical experiment (P,Q). In [5] it is used in the Bayesian setting to bound testing errors, through the following lemma.

Lemma 2.2. *For any set $\mathcal{P}$ of densities, and numbers $c, d > 0$, with ϕ ranging over all tests, and any $\alpha \in (0,1)$,*

$$\inf_\phi \sup_{P \in \mathcal{P}} (cP\phi + dQ(1-\phi)) \leq c^\alpha d^{1-\alpha} \rho_\alpha(\mathcal{P}, q).$$

In the intended applications the error probabilities $P\phi$ and $Q(1-\phi)$ are typically exponentially small, of the form $e^{-c\varepsilon^2}$ for $\varepsilon \to \infty$ and a positive constant c whose numerical value is not essential. Then there may also not be much loss in using affinities rather than tests, in particular in the symmetric case $c = d$, in view of the following lemma.

Lemma 2.3. *For any set $\mathcal{P}$ of probability densities, and numbers $c, d > 0$,*

$$\rho_{1/2}^2(\mathcal{P}, q) \leq \frac{c+d}{cd} \inf_{\phi} \sup_{P \in \mathcal{P}} (cP\phi + dQ(1-\phi)).$$

Proof. By the minimax theorem for testing the infimum over ϕ on the right side can be expressed as the supremum of $(c+d-\|cp-dq\|_1)/2$ over p ranging through the convex hull of $\mathcal{P}$. Furthermore, using the Cauchy-Schwarz inequality we can bound the square L_1-distance $\|cp-dq\|_1^2$ by $(c+d)^2 - 4cd\rho_{1/2}(p,q)$. Some algebra concludes the proof. □

The main tool in the following is the nonnegativeness of the Kullback-Leibler divergence (for probability densities), which is a well-known and immediate consequence of Jensen's inequality, and also of Lemma 2.1(v). For easy reference we state this fact in a slightly adapted form.

Lemma 2.4. *For a given, arbitrary nonnegative function v and a probability measure Π on a measurable space $\mathcal{P}$, we have for every probability density w relative to Π,*

$$\int (\log w) w \, d\Pi - \int (\log v) w \, d\Pi \geq -\log \int v \, d\Pi. \tag{3}$$

Equality is attained for $w \propto v$.

Proof. Were v a probability density, then the right side would be zero and the statement follows from the nonnegativity of the Kullback-Leibler information. A general function v can be normalized to a probability density by dividing by $\int v \, d\Pi$. Because $\int w \, d\Pi = 1$, this changes the left side by adding $\log \int v \, d\Pi$, which is independent of w and thus does not change the minimizing w. □

3. General result

The following theorem, due to [13], gives a bound on the concentration of the pseudo posterior Π_ρ, defined in (2).

Theorem 3.1. *For any numbers $\alpha \geq 0$, $\beta \in (0,1)$, $\gamma \geq 0$ and X distributed according to q, for $\rho = (\gamma\alpha + \beta)/(\gamma+1)$,*

$$\begin{aligned} \mathrm{E} \int R_\beta(p,q) \, d\Pi_\rho(p|\, X) \leq &-(\gamma+1) \log \int e^{-\rho KL(p,q)} \, d\Pi(p) \\ &+ \gamma \mathrm{E} \log \int \left(\frac{p}{q}\right)^\alpha (X) \, d\Pi(p). \end{aligned}$$

Proof. Applying (3) for a fixed observation X with $v(p) \propto (p/q)^\alpha(X)$, we find, for any probability density w relative to Π,

$$\int (\log w) w \, d\Pi - \alpha \int \left(\log \frac{p}{q}(X)\right) w(p) \, d\Pi(p) \geq -\log \int \left(\frac{p}{q}\right)^\alpha (X) \, d\Pi(p).$$

Applying (3) again, this time with $v(p) \propto (p/q)^\beta(X)/\rho_\beta(p,q)$, we find

$$\int (\log w) w \, d\Pi - \beta \int \left(\log \frac{p}{q}(X) \right) w(p) \, d\Pi(p) + \int \log \rho_\beta(p,q) \, w(p) \, d\Pi(p) \geq -\log c_\beta(X),$$

where $c_\beta(X) = \int (p/q)^\beta(X)/\rho_\beta(p,q) \, d\Pi(p)$ is the norming constant. We add the second inequality to γ times the first inequality. The resulting inequality can be reorganized into

$$\begin{aligned} &\int R_\beta(p,q) \, w(p) \, d\Pi(p) \\ &\quad \leq (\gamma+1) \int (\log w) w \, d\Pi - (\gamma\alpha + \beta) \int \left(\log \frac{p}{q}(X) \right) w(p) \, d\Pi(p) \\ &\qquad + \gamma \log \int \left(\frac{p}{q} \right)^\alpha (X) \, d\Pi(p) + \log c_\beta(X), \end{aligned} \tag{4}$$

If X is distributed according to the density q, then, by Jensen's inequality,

$$\mathrm{E} \log c_\beta(X) \leq \log \mathrm{E} c_\beta(X) = \log \int \frac{\mathrm{E}(p/q)^\beta(X)}{\rho_\beta(p,q)} \, d\Pi(p) \leq \log 1 = 0.$$

This shows that the last term on the right of (4) can be deleted after taking the expectation. The expectation $R := \mathrm{E}\gamma \log \int (p/q)^\alpha(X) \, d\Pi(p)$ of the second last term is copied to the bound given by the theorem. By Lemma 2.4 the remaining part of the right side (the difference of the first two terms) is minimized with respect to probability densities w, for fixed X, by $w(p) \propto p^\rho(X)$. For this minimizing function $w(p) \, d\Pi(p)$ in the left side becomes $d\Pi_\rho(p|\, X)$. It follows that

$$\begin{aligned} &\frac{1}{\gamma+1} \left[\mathrm{E} \int R_\beta(p,q) \, d\Pi_\rho(p|\, X) - R \right] \\ &\quad \leq \mathrm{E} \inf_w \left[\int (\log w) w \, d\Pi - \rho \int \left(\log \frac{p}{q}(X) \right) w(p) \, d\Pi(p) \right] \\ &\quad \leq \inf_w \left[\int (\log w) w \, d\Pi + \rho \int \mathrm{E} \left(\log \frac{q}{p}(X) \right) w(p) \, d\Pi(p) \right] \\ &\quad = \inf_w \left[\int (\log w) w \, d\Pi - \int \left(\log e^{-\rho KL(p,q)} \right) w(p) \, d\Pi(p) \right] \\ &\quad = -\log \int e^{-\rho KL(p,q)} \, d\Pi(p). \end{aligned}$$

Here the last step follows again by Lemma 2.4. □

The Renyi divergence $R_\beta(p,q)$ is nonnegative and vanishes for $p = q$. Hence the left side of the theorem can be viewed as a measure for the concentration of the pseudo posterior distribution near q. The easiest interpretation is obtained by bounding the Renyi divergence below by the Hellinger distance, e.g. for $\beta = 1/2$ twice the left side of Theorem 3.1 is an upper bound on $\mathrm{E} \int h^2(p,q) \, d\Pi_\rho(p|\, X)$, as $2R_{1/2} \geq h^2$.

The first term on the right of the theorem is a measure of concentration of the prior Π near q. As $e^{-\rho KL(p,q)} \leq 1$ for all p and Π is a probability measure, this term is always nonnegative; it is near zero if $KL(p,q) \approx 0$ with high prior probability. An

explicit bound, following from Markov's inequality $\mathrm{E}e^{-\rho Z} \geq e^{-\rho z}\mathrm{P}(Z < z)$, valid for any variable Z and any z, is

$$-\log \int e^{-\rho KL(p,q)}\, d\Pi(p) \leq \varepsilon^2\rho - \log \Pi(p\colon KL(p,q) < \varepsilon^2).$$

The right side is bounded by $\varepsilon^2(\rho + c)$ if

$$\Pi\big(p\colon KL(p,q) < \varepsilon^2\big) \geq e^{-c\varepsilon^2}. \tag{5}$$

This is a version of the prior mass condition in [2] or [3] stripped from any reference to a sampling model. The condition requires that the prior sufficiently charges Kullback-Leibler neighbourhoods of q, and in some form is necessary for sufficient posterior concentration near q.

The downside of the theorem is the second term on its right side. By Jensen's inequality,

$$\mathrm{E}\log \int \left(\frac{p}{q}\right)^{\alpha}(X)\, d\Pi(p) \leq \log \int \rho_\alpha(p,q)\, d\Pi(p). \tag{6}$$

For $\alpha \leq 1$, the Hellinger transform $\rho_\alpha(p,q)$ is bounded by 1, and hence the right side is bounded above by $\log 1 = 0$. For $\alpha > 1$, the inequality is still valid, but the right side may not even be finite.

Therefore, for $\alpha \leq 1$ the second term of the upper bound can be omitted and the theorem is very satisfying; for $\alpha > 1$ additional arguments are necessary. Closer inspection shows that the case $\alpha \leq 1$ covers the pseudo posteriors with $\rho < 1$, but unfortunately excludes the true posterior ($\rho = 1$) and pseudo posteriors with $\rho > 1$. The parameters are related by

$$\rho = \frac{\gamma\alpha + \beta}{\gamma + 1}.$$

For fixed $\alpha \geq \beta$ the parameter ρ increases from β to α as γ increases from 0 to ∞; for $\alpha < \beta$ it decreases from β to α. Any choice $\beta < 1$ requires to choose $\alpha > 1$ to reach $\rho = 1$ for some finite γ.

On the other hand, any $\rho < 1$ is possible. Combined with the preceding observations this yields the following corollary.

Corollary 3.1. *If (5) holds for given $c, \varepsilon > 0$, then for any $\rho < 1$,*

$$\mathrm{E}\int h^2(p,q)\, d\Pi_\rho(p|\, X) \leq \frac{\varepsilon^2(\rho + c)}{((1-\rho)\wedge\rho)}.$$

Proof. We use Theorem 3.1 with $\beta = 1/2$, so that its left side is an upper bound on twice the left side of the lemma, with the first term on its right side bounded using the prior mass condition (5) as indicated, and with a value of α smaller than 1, so that the second term on its right side is bounded above by 0.

For $0 < \rho < 1/2$ we choose $\alpha = 0$ and $\gamma + 1 = 1/(2\rho)$; for $\rho = 1/2$ we choose $\gamma = 0$; and for $1/2 < \rho < 1$ we choose $\alpha = 1$ and $\gamma = (\rho - 1/2)/(1-\rho)$, giving $\gamma + 1 = 1/(2(1-\rho))$. □

For $\alpha > 1$ the second term in the bound must be analyzed separately. This difficulty reflects the finding that posterior contraction cannot be ensured by sufficient prior mass in a neighbourhood of the true density alone, but the full model, or

the spread of the posterior over the model, must be taken into account. Various approaches have made this precise. Conditions that imply the existence of good tests of q versus elements of $\mathcal{P}$ are one possibility. As shown by [8] and [1] bounds on the metric entropy of (subsets of) $\mathcal{P}$ ensure existence of suitable tests. Tests are related to affinities, as shown in Lemma 2.2. The next theorem shows that affinities may also be used to analyze the additional term.

Following [5] for $\beta \in (0,1)$ and an arbitrary metric d on $\mathcal{P}$ define the *covering number for testing* $N_{t,\beta}(\varepsilon, \mathcal{P}, d)$ (for $\varepsilon > 0$) as the minimal number of sets $B_1, \ldots, B_N$ needed to cover $\{p \in \mathcal{P}: \varepsilon \le d(p,q) < 2\varepsilon\}$ and such that

$$R_\beta(B_i, q) \ge \frac{\varepsilon^2}{4}, \qquad i = 1, \ldots, N.$$

Theorem 3.2. *Let $\mathcal{P} = \cup_{k\in K}\mathcal{P}_k$ be a countable partition of $\mathcal{P}$ such that $N_{t,\beta}(\varepsilon, \mathcal{P}_k, d) \le N_k(\varepsilon)$ for every $\varepsilon \ge \varepsilon_0 > 0$, for nonincreasing functions $N_k: (0,\infty) \to \mathbb{R}$. If (5) holds, then for any $0 < \delta < \beta < 1$, any $\varepsilon > \varepsilon_0$, and for X distributed according to q,*

$$\frac{1}{16}\mathrm{E}\int_{p:KL(p,q)\ge\varepsilon^2} d^2(p,q)\, d\Pi(p|\, X) \le \varepsilon^2\left[1 + \frac{(1+c)\beta(1-\delta)}{\beta-\delta}\right] + \frac{1-\beta}{\beta-\delta}\log\left[2 + 4\varepsilon_0^{-2}\sum_{k\in K} N_k(\varepsilon)\Pi(\mathcal{P}_k)^\delta\right].$$

Proof. Let $\mathcal{P}_{0,1,*} = \{p \in \mathcal{P}: KL(p,q) < \varepsilon^2\}$, $\mathcal{P}_{0,2,*} = \{p \in \mathcal{P}: d(p,q) < \varepsilon\}$, and for $i = 1, 2, \ldots$ and $k \in K$ let $\mathcal{P}_{i,1,k}, \ldots, \mathcal{P}_{i,N_{i,k},k}$ be a minimal cover of the set $\{p \in \mathcal{P}_k: i\varepsilon \le d(p,q) < (i+1)\varepsilon\}$ by sets such that $R_\beta(\mathcal{P}_{i,j,k}, q) \ge i^2\varepsilon^2/4$, for every (j,k). By the definition of the covering numbers for testing we can choose $N_{i,k} \le N_{t,\beta}(i\varepsilon, \mathcal{P}_k, d) \le N_k(i\varepsilon) \le N_k(\varepsilon)$ for $i \ge 1$. Make the sets $\mathcal{P}_{i,j,k}$ disjoint by sequentially omitting previous sets, thus giving a partition $\{\mathcal{P}_{i,j,k}\}$ of $\mathcal{P}$, indexed by $M := \{(i,j,k): i = 1,2,\ldots; j = 1,\ldots,N_{i,k}; k \in K\} \cup \{(0,1,*),(0,2,*)\}$.

If $p \in \mathcal{P}_{i,j,k}$ for $i \ge 1$, then $d^2(p,q) \le (i+1)^2\varepsilon^2 \le 16R_\beta(\mathcal{P}_{i,j,k}, q)$. Consequently

$$\frac{1}{16}\int_{p\notin\mathcal{P}_{0,1,*}\cup\mathcal{P}_{0,2,*}} d^2(p,q)\, d\Pi(p|\, X) \le \sum_{(i,j,k)} R_\beta(\mathcal{P}_{i,j,k}, q)\, \Pi(\mathcal{P}_{i,j,k}|\, X). \tag{7}$$

In the right side we can replace $\mathcal{P}_{i,j,k}$ in $R_\beta(\mathcal{P}_{i,j,k}, q)$, in view of the latter's definition as an infimum, by any $p_{i,j,k}$ in the convex hull of $\mathcal{P}_{i,j,k}$.

View the numbers $(\Pi(\mathcal{P}_{i,j,k}): (i,j,k) \in M)$ as a prior on the model $(p_{i,j,k}: (i,j,k) \in M)$ consisting of the densities $p_{i,j,k}$ defined by

$$p_{i,j,k} = \int_{\mathcal{P}_{i,j,k}} p\, \frac{d\Pi(p)}{\Pi(\mathcal{P}_{i,j,k})}.$$

The corresponding posterior gives the posterior probabilities of the densities $p_{i,j,k}$ and can be identified with the collection of numbers

$$\frac{p_{i,j,k}(X)\Pi(\mathcal{P}_{i,j,k})}{\sum_{(i,j,k)} p_{i,j,k}(X)\Pi(\mathcal{P}_{i,j,k})} = \frac{\int_{\mathcal{P}_{i,j,k}} p(X)\, d\Pi(p)}{\int p(X)\, d\Pi(p)} = \Pi(\mathcal{P}_{i,j,k}|\, X).$$

In other words, the posterior in this "discretized setting" is the collection $(\Pi(\mathcal{P}_{i,j,k}|\, X): (i,j,k) \in K)$ of posterior probabilities of the partitioning sets in the original setting.

By Theorem 3.1 applied with $\rho = 1$, the given β, and α and γ satisfying $\gamma\alpha+\beta = \gamma+1$, the expected value of the right side of (7) is bounded above by

$$-(\gamma+1)\log \sum_{(i,j,k)} e^{-KL(p_{i,j,k},q)}\Pi(\mathcal{P}_{i,j,k}) + \gamma \mathrm{E}\log \sum_{(i,j,k)} \left(\frac{p_{i,j,k}}{q}\right)^{\alpha}(X)\Pi(\mathcal{P}_{i,j,k}).$$

The first term becomes bigger if we leave off all terms of the sum except the $(0,1,*)$-term, which is

$$-(\gamma+1)\log\bigl[e^{-KL(p_{0,1,*},q)}\Pi(\mathcal{P}_{0,1,*})\bigr] \le (\gamma+1)(1+c)\varepsilon^2,$$

in view of (5). By the subadditivity of the map $x \mapsto x^{\delta}$, for $\delta \le 1$, the second term is bounded by

$$\frac{\gamma}{\delta}\mathrm{E}\log \sum_{(i,j,k)} \left(\frac{p_{i,j,k}}{q}\right)^{\alpha\delta}(X)\Pi(\mathcal{P}_{i,j,k})^{\delta} \le \frac{\gamma}{\delta}\log \sum_{(i,j,k)} \rho_{\alpha\delta}(p_{i,j,k},q)\Pi(\mathcal{P}_{i,j,k})^{\delta},$$

by Jensen's inequality and concavity of the logarithm. We choose $\alpha\delta = \beta < 1$ and then have that $\rho_{\alpha\delta}(p_{i,j,k},q)$ is bounded by 1 for any (i,j,k) and equal to $\rho_{\beta}(p_{i,j,k},q) = e^{-R_{\beta}(p_{i,j,k},q)} \le e^{-i^2\varepsilon^2/4}$, for the remaining terms (i,j,k). Since $\mathcal{P}_{i,j,k} \subset \mathcal{P}_k$, and there are at most $N_k(\varepsilon)$ indices j for given (i,k), the series is bounded by $2+\sum_{i\ge 1}\sum_k N_k(\varepsilon)e^{-i^2\varepsilon^2/4}\Pi(\mathcal{P}_k)^{\delta} = 2+\sum_k N_k(\varepsilon)\Pi(\mathcal{P}_k)^{\delta}/(e^{\varepsilon^2/4}-1)$. Here $e^{\varepsilon^2/4}-1 \ge \varepsilon^2/4 \ge \varepsilon_0^2/4$.

For the given choices of parameters we have $\gamma/\delta = (1-\beta)/(\beta-\delta)$ and $\gamma+1 = \beta(1-\delta)/(\beta-\delta)$. This yields the bound as in the theorem. □

The partition $\mathcal{P} = \cup_k \mathcal{P}_k$ in the theorem allows to trade off the complexity of submodels $\mathcal{P}_k$ versus their prior masses, similarly as in [4]. For simplicity in the following we restrict to a partition in one set (no partition).

The theorem makes no assumption on the sampling model for the observation X, and uses a distance on the full data model. Notwithstanding the notation, it will typically be applied with a *large* ε. The factor 2 inside the logarithm will then be negligible and a rate ε^2 is attained if $\sum_k N_k(\varepsilon)\Pi(\mathcal{P}_k)^{\delta} \lesssim e^{\varepsilon^2}$.

From the convexity of Hellinger balls and Lemma 2.1(viii), it can be seen that for d the Hellinger distance the covering numbers for testing are dominated by the more usual *local covering numbers* or *Le Cam dimension*:

$$N_{t,\beta}(\varepsilon,\mathcal{P},h) \le N(\varepsilon b, \{p\in\mathcal{P}\colon \varepsilon < h(p,q) \le 2\varepsilon\}, h),$$

where $b = 1-(\beta\wedge(1-\beta))^{-1/2}/2$ and $N(\varepsilon,\mathcal{P},d)$ is the minimal number of balls of radius ε needed to cover $\mathcal{P}$ (cf. [10], [5], page 642; for $\beta = 1/2$ we can use $b = 1/4$). This observation allows to deduce a result that is analogous to the main result of [2].

Corollary 3.2. *Suppose that* $N(\varepsilon/4, \{p\in\mathcal{P}\colon \varepsilon \le d(p,q) < 2\varepsilon\}, h) \le N(\varepsilon)$ *for every* $\varepsilon > \varepsilon_0$ *and a nonincreasing function* $N\colon (0,\infty) \to \mathbb{R}$. *If (5) holds, then, for* X *distributed according to* q *and every* $\varepsilon > \varepsilon_0$,

$$\frac{1}{16}\mathrm{E}\int h^2(p,q)\,d\Pi(p|\,X) \le \varepsilon^2(3+c) + \log N(\varepsilon) + \log_+(4/\varepsilon_0^2) + \log 3.$$

Proof. We apply the theorem with $d = h$, $\beta = 1/2$ and a partition in a single set. We bound $\Pi(\mathcal{P})^\delta$ by 1, and next let $\delta \downarrow 0$. Then the parameter in square brackets tends to $2 + c$, and the parameter in front of the logarithm tends to $(1 - \beta)/\beta = 1/2$. Because $h^2 \leq KL$, the "missing part" of the integral, over the set $\{p: KL(p, q) < \varepsilon^2\}$, is bounded by ε^2, raising $2 + c$ to $3 + c$. Finally we simplify using the inequalities $\log(2 + x) \leq \log 3 + \log_+ x$ and $\log_+(xy) \leq \log_+ x + \log_+ y$, for any $x, y > 0$. □

An alternative method, evoked in [13], to estimate the remainder term in Theorem 3.1 for $\alpha > 1$ is to cover the support of the prior by (upper) brackets. For any partition $\mathcal{P} = \cup_{j=1}^N \mathcal{P}_j$, by subadditivity of the map $x \mapsto x^{1/\alpha}$, for $\alpha > 1$, and Jensen's inequality,

$$\begin{aligned} \mathrm{E} \log \int \left(\frac{p}{q}\right)^\alpha (X)\, d\Pi(p) &\leq \alpha\, \mathrm{E} \log \sum_{j=1}^N \left(\sup_{p \in \mathcal{P}_j} \frac{p}{q}\right)(X)\, \Pi(\mathcal{P}_j)^{1/\alpha} \\ &\leq \alpha \log \sum_{j=1}^N \left(\int \sup_{p \in \mathcal{P}_j} p\, d\mu\right) \Pi(\mathcal{P}_j)^{1/\alpha}. \end{aligned}$$

A crude bound on the sum in the right side is $N \max_j \int \sup_{p \in \mathcal{P}_j} p\, d\mu$. Because $p \in \mathcal{P}_j$ are probability densities, the integral will be bigger than 1. By constructing the partition from a minimal set $[l_1, u_1], \ldots, [l_N, u_N]$ of ε^2-brackets in $L_1(\mu)$ that covers $\mathcal{P}$, the overshoot is at most ε^2, and the preceding display can be bounded by

$$\alpha \log N_{[]}(\varepsilon^2, \mathcal{P}, L_1(\mu)) + \alpha\varepsilon^2.$$

Unfortunately, this approach does not appear to yield the "correct" rate in general. For this we would like to see the entropy $\log N(\varepsilon, \mathcal{P}, d)$ at ε, and not at ε^2, in the bound, probably for another metric d than the $L_1(\mu)$-metric. One might try to compensate this by taking also the prior masses into account; see e.g. [6] for results in this direction.

In the following section we use empirical process methods to improve the bracketing approach in the case of i.i.d. observations.

4. Independent experiments

If the observation is a random sample $X_1, \ldots, X_n$ of size n, then we apply the preceding with p and q product densities. The Hellinger affinity is multiplicative and the Renyi divergence and Kullback-Leibler divergence are additive relative to independent observations. For collections of measures we have defined these quantities by taking the supremum or infimum over the convex hull. This destroys exact multiplicativity or additivity, but sub-multiplicativity and super- or sub- additivity are retained.

Given sets $\mathcal{P}_i$ of densities relative to dominating measures μ_i on measurable spaces $(\mathcal{X}_i, \mathcal{A}_i)$, let $\mathcal{P}_1 \times \mathcal{P}_2$ denote the set of all densities $(x_1, x_2) \mapsto p_1(x_1)p_2(x_2)$ relative to $\mu_1 \otimes \mu_2$.

Lemma 4.1. *For any sets $\mathcal{P}_1, \mathcal{P}_2$ of probability densities and probability densities q_1, q_2 and any $\alpha \in (0, 1)$,*

$$\begin{aligned} \rho_\alpha(\mathcal{P}_1 \times \mathcal{P}_2, q_1 \times q_2) &\leq \rho_\alpha(\mathcal{P}_1, q_1)\rho_\alpha(\mathcal{P}_2, q_2), \\ R_\alpha(\mathcal{P}_1 \times \mathcal{P}_2, q_1 \times q_2) &\geq R_\alpha(\mathcal{P}_1, q_1) + R_\alpha(\mathcal{P}_2, q_2), \\ KL(\mathcal{P}_1 \times \mathcal{P}_2, q_1 \times q_2) &= KL(\mathcal{P}_1, q_1) + KL(\mathcal{P}_2, q_2). \end{aligned}$$

Proof. The first inequality is due to Le Cam (also see [5], p. 866, or [13]). It follows from writing $\rho_\alpha(\int p_1 \times p_2 \, d\Pi(p_1, p_2), q_1 \times q_2)$ for a given probability measure Π in the form

$$\int \left[\int \left(\int p_1(x_1) \frac{\int p_2(x_2)\, d\Pi_{2|1}(p_2|\, p_1)}{\int p_2(x_2)\, d\Pi_2(p_2)} \, d\Pi_1(p_1) \right)^\alpha q_1(x_1)^{1-\alpha} \, d\mu_1(x_1) \right] \times \left(\int p_2(x_2)\, d\Pi_2(p_2) \right)^\alpha q_2(x_2)^{1-\alpha} \, d\mu_2(x_2).$$

Here Π_i are the marginal distributions of Π and $\Pi_{2|1}$ is a conditional distribution (in the sense that $d\Pi_{2|1}(p_2|\, p_1)\, d\Pi_1(p_1) = d\Pi(p_1, p_2)$; no regularity condition on existence of a conditional is necessary). The term within square brackets is bounded above by $\rho_\alpha(\mathcal{P}_1, q_1)$. Next the remaining integral is bounded above by $\rho_\alpha(\mathcal{P}_2, q_2)$. The second inequality is an immediate consequence.

To prove the third we first note the Kullback-Leibler divergence is convex (in both its arguments), whence the convex hull in the definition of $KL(\mathcal{P}, q)$ is unnecesaary: this is equal to $\sup_{p\in\mathcal{P}} KL(p, q)$. The assertion then follows from the additivity: $KL(p_1 \times p_2, q_1 \times q_2) = KL(p_1, q_1) + KL(p_2, q_2)$. □

Consider an application of Theorem 3.2 to the case of i.i.d. observations from a density q and a prior Π on a model $\mathcal{P}$ for one observation. Thus the model $\mathcal{P}$ in Theorem 3.2 is the model $\mathcal{P}^n = \{p^{\times n} : p \in \mathcal{P}\}$ in the present set-up. We replace ε in Theorem 3.2 by $\sqrt{n}\varepsilon$ and the metric d on $\mathcal{P}^n$ by $\sqrt{n}h$ for h the Hellinger distance on the model $\mathcal{P}$ for one observation. The prior mass condition (5) becomes

$$\Pi\big(p : KL(p, q) < \varepsilon^2\big) \geq e^{-cn\varepsilon^2}. \tag{8}$$

Corollary 4.1. *Suppose that $N(\varepsilon/4, \{p \in \mathcal{P} : \varepsilon < d(p, q) < 2\varepsilon\}, h) \leq N(\varepsilon)$ for every $\varepsilon > 0$ and a nonincreasing function $N : (0, \infty) \to \mathbb{R}$. If (8) holds, then, for $X_1, \ldots, X_n$ an i.i.d. sample from q and $\varepsilon \geq 1/\sqrt{n}$,*

$$\frac{1}{16} \mathrm{E} \int h^2(p, q)\, d\Pi(p|\, X_1, \ldots, X_n) \leq \varepsilon^2(3 + c) + \frac{1}{n} \log N(\varepsilon) + \frac{1}{n} \log 12.$$

Proof. This follows from Theorem 3.2 upon making the substitutions as explained, and using the inequality $N_{t,\beta}(\sqrt{n}\varepsilon, \mathcal{P}^n, d) \leq N_{t,\beta}(\varepsilon, \mathcal{P}, h)$. □

For $\log N(\varepsilon_n) \asymp n\varepsilon_n^2$ the bound is of the order ε_n^2. This is the "correct" expression of the rate in the complexity of the model (cf. [8], [1]).

We have not been able to bound the concentration of pseudo posterior distributions with $\rho > 1$ by similar arguments. It seems that stronger control of the model than just covering numbers are needed. For maximum likelihood estimators (the case $\rho = \infty$) a basic result due to [12] is in terms of the *bracketing integral*

$$J_{[]}(\delta, \mathcal{P}, h) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{P}, h)}\, d\varepsilon,$$

where $N_{[]}(\varepsilon, \mathcal{P}, h)$ is the minimal number of ε-brackets relative to the Hellinger distance needed to cover $\mathcal{P}$ (see Definition 2.1.6 of [11]). The maximum likelihood estimator converges at rate ε_n equal to the minimal solution to

$$J_{[]}(\varepsilon, \mathcal{P}, h) \leq \sqrt{n}\varepsilon^2. \tag{9}$$

(See [12], or [11], Section 3.4.1.) If $J_{[]}(\varepsilon, \mathcal{P}, h) \asymp \varepsilon(\log N_{[]}(\varepsilon, \mathcal{P}, h))^{1/2}$, which is the case if the bracketing entropy varies regularly, then this reduces to $\log N_{[]}(\varepsilon, \mathcal{P}, h) \lesssim n\varepsilon^2$, which can be compared to the rate obtained in Corollary 4.1.

The pseudo posterior contracts at the same rate.

Theorem 4.1. *If ε satisfies (8) and (9), then, for $X_1, \ldots, X_n$ an i.i.d. sample from q and any $\rho > 0$,*

$$\mathrm{E} \int h^2(p, q)\, d\Pi_\rho(p|\, X_1, \ldots, X_n) \lesssim \varepsilon^2.$$

Proof. We apply Theorem 3.1 to the product densities, with the substitutions explained before the statement of Corollary 4.1. It suffices to bound the last term on the right side of Theorem 3.1, for some $\alpha > \rho$, so that there exists $\gamma \in (0, \infty)$ with $\rho = (\alpha\gamma + \beta)/(\gamma + 1)$ for some $\beta \in (0, 1)$ (e.g. $\beta = 1/2$), whence $R_\beta \gtrsim h^2$.

Let $\mathbb{G}_n$ be the empirical process of $X_1, \ldots, X_n$, and for $\tau < 0$ define $\log_\tau x = (\log x) \vee \tau$.

By Lemmas 4 and 5 in [12] there exists $\tau < 0$ such that $Q \log_\tau(p/q) \le -c\, h^2(p, q)$ and $\| \log_\tau(p/q)/2 \|_{Q,B} \le d\, h(p, Q)$, for positive constants c, d that depend on τ only, where $\|\cdot\|_{Q,B}$ is the "Bernstein norm" defined in [11], page 324. Furthermore, following the approach of Theorem 3.4.4 of [11] it can be shown that there exist a constant e, which also depends on τ only, such that $\| \log_\tau(p_2/q) - \log_\tau(p_2/q) \|_{Q,B} \le e\, h(p_1, p_2)$, for every pair of functions with $p_1 \le p_2$. These facts imply, by extension of Lemma 3.4.3 in [11] to higher moments, that, for any $\delta > 0$,

$$\mathrm{E} \sup_{h(p,q) \le \delta} \big(\mathbb{G}_n \log_\tau(p/q)\big)_+^4 \lesssim J_{[]}^4(\delta, \mathcal{P}, h) \left(1 + \frac{J_{[]}(\delta, \mathcal{P}, h)}{\sqrt{n}\delta^2}\right)^4. \tag{10}$$

Since $\delta \mapsto J_{[]}(\delta, \mathcal{P}, h)$ is the area under a decreasing, nonnegative function, the function $\delta \mapsto J_{[]}(\delta, \mathcal{P}, h)/\delta$ is decreasing. First this shows that $J_{[]}(C\delta, \mathcal{P}, h) \le C J_{[]}(\delta, \mathcal{P}, h)$, for every $C > 1$. Second the function $\delta \mapsto J_{[]}(\delta, \mathcal{P}, h)/\delta^2$ is also decreasing, implying that (9) holds for any ε bigger than its minimal solution. Therefore for δ bigger than this minimal solution the quotient inside the brackets in (10) is bounded by one and the right side can be simplified to $J_{[]}^4(\delta, \mathcal{P}, h)$.

For integers $i \ge 1$ define $\mathcal{P}_i = \{p \in \mathcal{P} \colon 2^{i-1}\varepsilon \le h(p, q) < 2^i\varepsilon\}$; also set $\mathcal{P}_0 = \{p \in \mathcal{P} \colon h(p, q) < \varepsilon\}$. Then $Q \log_\tau(p/q)$ is bounded above by $-ch^2(p, q) \le -c2^{2i-2}\varepsilon^2$ if $p \in \mathcal{P}_i$ and $i \ge 1$, and is nonpositive for $p \in \mathcal{P}_0$. Because $\log x \le \log_\tau x$ for every $x > 0$,

$$\begin{aligned}
&\frac{1}{\alpha n} \mathrm{E} \log \int \left(\frac{p^{\times n}}{q^{\times n}}\right)^\alpha (X)\, d\Pi(p) \\
&\quad \le \frac{1}{n} \mathrm{E} \sup_{p \in \mathcal{P}} \log_\tau \frac{p^{\times n}}{q^{\times n}}(X) \\
&\quad \le \mathrm{E} \sup_{p \in \mathcal{P}_0} \frac{1}{\sqrt{n}} \left(\mathbb{G}_n \log_\tau \frac{p}{q}\right)_+ + \mathrm{E} \sup_{i \ge 1} \left(\sup_{p \in \mathcal{P}_i} \frac{1}{\sqrt{n}} \mathbb{G}_n \log_\tau \frac{p}{q} - c2^{2i-2}\varepsilon^2 \right)_+ .
\end{aligned}$$

By (10) the first expectation on the right is bounded above by a multiple of $n^{-1/2} J_{[]}(\varepsilon, \mathcal{P}, h) \le \varepsilon^2$. To bound the second term we apply Markov's inequality to see that, for $x > 0$,

$$\begin{aligned}
\mathrm{P}\left(\sup_{p \in \mathcal{P}_i} \frac{1}{\sqrt{n}} \mathbb{G}_n \log_\tau \frac{p}{q} - c2^{2i-2}\varepsilon^2 > x \right) &\le \frac{\mathrm{E}(\sup_{p \in \mathcal{P}_i} \mathbb{G}_n \log_\tau(p/q))_+^4}{n^2(x + c2^{2i-2}\varepsilon^2)^4} \\
&\lesssim \frac{J_{[]}^4(2^i\varepsilon, \mathcal{P}, h)}{n^2(x + c2^{2i-2}\varepsilon^2)^4}.
\end{aligned}$$

Here $J_{[]}(2^i\varepsilon, \mathcal{P}, h) \leq 2^i J_{[]}(\varepsilon, \mathcal{P}, h) \leq 2^i\sqrt{n}\varepsilon^2$, for ε satisfying (9). It follows that the second expectation in the far right side of the second last display is bounded above by

$$\int_0^\infty \sum_{i=1}^\infty \frac{2^{4i}\varepsilon^8}{(x + c2^{2i-2}\varepsilon^2)^4}\, dx = \varepsilon^2 \sum_{i=1}^\infty \frac{2^{-2i}2^6}{3c^3} \lesssim \varepsilon^2.$$

This concludes the proof. □

5. Misspecification

The right side of Theorem 3.1 can be small only if $KL(p, q)$ is close to zero with sufficient prior mass (for $p \sim \Pi$). Therefore, the theorem does not cover the case that the density q of the observation is not close to the support of the prior. To remedy this we adapt the derivation as follows. Let q still be the true density of the observation and let $\tilde{q}$ be another density, later taken to the "projection" of q on the model.

Theorem 5.1. *For any numbers $\alpha \geq 0$, $\beta \in (0, 1)$, $\gamma \geq 0$ and X distributed according to q, for $\rho = (\gamma\alpha + \beta)/(\gamma + 1)$,*

$$\mathrm{E} \int R_\beta(pq/\tilde{q}, q)\, d\Pi_\rho(p|\, X) \leq -(\gamma + 1) \log \int e^{-\rho(KL(p,q) - KL(\tilde{q},q))}\, d\Pi(p) + \gamma \mathrm{E} \log \int \left(\frac{p}{\tilde{q}}\right)^\alpha (X)\, d\Pi(p).$$

Proof. We follow the same steps as in the proof of Theorem 3.1, except that we make the choices, first $v(p) \propto (p/\tilde{q})^\alpha(X)$ and second $v(p) \propto (p/\tilde{q})^\beta(X)/\rho_\beta(pq/\tilde{q}, q)$. □

The bound of the theorem is true for any $\tilde{q}$. However, it is clear that the first term on the right can be small only if the prior puts sufficient mass on densities p such that $KL(p, q) - KL(\tilde{q}, q) = Q \log \tilde{q}/p$ is close to zero, i.e. on densities p close to $\tilde{q}$. Furthermore, the theorem is useless unless $R_\beta(pq/\tilde{q}, q)$ is nonnegative. Because $pq/\tilde{q}$ is not a probability density, this is not guaranteed, not even when $\beta \in (0, 1)$. This is illustrated in Figure 1, taken from [5]. The Renyi divergence $R_\beta(pq/\tilde{q}, q)$ is positive if and only if the Hellinger affinity $\rho_\beta(pq/\tilde{q}, q)$ is bounded above by 1. As a function of β the Hellinger affinity is convex with right limit $Q(p > 0)$ at $\beta = 0$ and left limit $\int_{q>0} pq/\tilde{q}\, d\nu$ at $\beta = 1$. If the latter limit is strictly bigger than 1, then there are two cases:

1. The right derivative at $\beta = 0$ is negative; then there exists $\beta > 0$ for which $\rho_\beta(pq/\tilde{q}, q) \leq 1$.
2. The right derivative at $\beta = 0$ is positive; then $\rho_\beta(pq/\tilde{q}, q) \geq Q(p > 0)$, which is typically one, throughout $(0, 1)$.

By Lemma 2.1, if the distributions are absolutely continuous, this right derivative is equal to $-KL(pq/\tilde{q}, q) = KL(\tilde{q}, q) - KL(p, q)$. We conclude that $R_\beta(pq/\tilde{q}, q)$ will be positive for some β for a set $\mathcal{P}$ of p only if $\tilde{q}$ is chosen to minimize the Kullback-Leibler divergence $p \mapsto KL(p, q)$ over $\mathcal{P}$.

This argument is made in [5] in a testing context, accompanied with examples where $\rho_\beta(pq/\tilde{q}, q) < 1$ for a sufficiently small $\beta > 0$, uniformly in densities p in the support of the prior, and where $R_\beta(pq/\tilde{q}, q)$ is bounded below by a natural distance. It would be interesting to investigate similar consequences of Theorem 5.1.

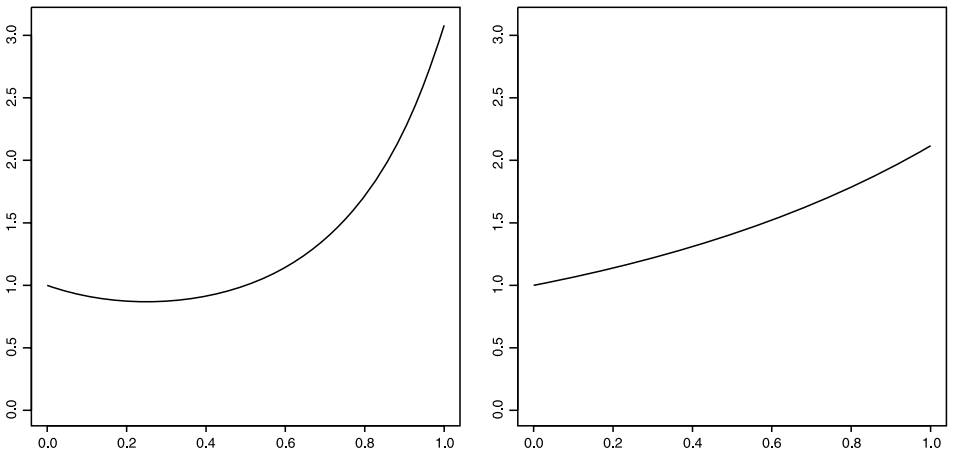

FIG 1. *The Hellinger transforms $\beta \mapsto \rho_\beta(p,q)$, for $Q = N(0,2)$ and P the measure defined by $dP = (dN(3/2,1)/dN(0,1))\,dQ$ (left) and $dP = (dN(3/2,1)/dN(1,1))\,dQ$ (right). Intercepts with the vertical axis at the right and left of the graphs equal $\|Q\| = 1 = Q(p > 0)$ and $\|P\| = P(q > 0)$ respectively. The slope at 0 equals $-KL(p,q)$, and has different sign in the two cases.*

References

[1] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **65** 181–237. URL http://dx.doi.org/10.1007/BF00532480

[2] GHOSAL, S., GHOSH, J. K. AND VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. URL http://dx.doi.org/10.1214/aos/1016218228

[3] GHOSAL, S. AND VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. URL http://dx.doi.org/10.1214/009053606000001172

[4] GHOSAL, S. AND VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. URL http://dx.doi.org/10.1214/009053606000001271

[5] KLEIJN, B. J. K. AND VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. URL http://dx.doi.org/10.1214/009053606000000029

[6] KRUIJER, W. (2008). Convergence rates in nonparametric Bayesian density estimation. Ph.D. thesis, VU University Amsterdam.

[7] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics*. Springer-Verlag, New York.

[8] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

[9] SCHWARTZ, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **4** 10–26.

[10] VAN DER VAART, A. (2002). The statistical work of Lucien Le Cam. *Ann. Statist.* **30** 631–682. Dedicated to the memory of Lucien Le Cam. URL http://dx.doi.org/10.1214/aos/1028674836

[11] VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer-Verlag, New York. With applications to statistics.

[12] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362. URL http://dx.doi.org/10.1214/aos/1176324524

[13] Zhang, T. (2006). From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. URL http://dx.doi.org/10.1214/009053606000000704